Practical Implementations of Speaker-Adaptive Training

†*Spyros Matsoukas, Rich Schwartz, Hubert Jin, Long Nguyen*

BBN Systems and Technologies
70 Fawcett Street, Cambridge, MA 02138
†Northeastern University, Boston, MA

ABSTRACT

Speaker Adaptive Training (SAT) has been shown to achieve significant word error reductions relative to the common Speaker Independent (SI) training paradigm, but its high requirements in disk I/O and space make it impractical for training on more than a couple hundred speakers. In the 1996 Hub-4 evaluation, the 38 hours of broadcast news training data consist of approximately 2000 speakers, half of them having less than 20 seconds of speech. In this paper we propose three implementations of SAT that are practical for training sets with a few thousands of speakers. First we present a two-pass SAT procedure that is mathematically equivalent to the original SAT method, with significantly reduced requirements in disk space, but essentially double the training time. Then we describe the Inverse Transform SAT (ITSAT) and the Least Squares SAT (LSSAT), two approximations to the SAT parameter estimation with time and space requirements that match those of common SI training. We show that the ITSAT method suffers only 1% degradation relative to the original SAT method.

1. INTRODUCTION

The ultimate goal of automatic speech recognition has always been to achieve human-level performance in real life applications. What makes this goal so hard to achieve is that in real life situations speech is usually the conglomaration of many diverse components, such as background effects (noise, other speakers, music), channel characteristics (room acoustics, telephone), and inter-speaker variability. This fact implies that any system that wants to perform well in real life applications has to be trained on speech that includes all the above diverse characteristics. Most importantly, it has to be able to estimate accurately the parameters of the acoustic model by annihilating as much variability as possible.

Speaker-Adaptive Training (SAT) [1] is a method of estimating the parameters of continuous density HMMs for speaker-independent (SI) continuous speech recognition in a way that integrates speaker adaptation [2] in the common SI training paradigm. This method works particularly well on training sets that contain speech from a large population of speakers as well as speech with varying recording and/or channel conditions. Unfortunately, as we will see in the following section, the original SAT implementation requires significant amount of disk space for each speaker in the training. The 1996 Hub-4 Broadcast News evaluation challenges SAT with a training set that contains approximately 2000 speakers in many different channel conditions, all in just 38 hours of speech. Because of practical limitations, the original SAT method cannot handle more than a few hundreds of speakers, so one way to use SAT on large training populations is to cluster the speakers down to a practical number. But this method contradicts the motivation of SAT, which is to model separately the speaker specific variation from the other phonetically relevant variation of the speech signal, and obtain reduced variance acoustic models.

In this paper, we propose three alternative implementations of SAT that are practical for training on thousands of speakers, without requiring speaker clustering.

2. ORIGINAL SAT METHOD

Before we present the practical implementations of SAT, it would be useful to describe briefly the original SAT parameter estimation.

As in [1], we assume a set of continuous density HMM triphone models with N states, where the j-th state observation density is assumed to be a mixture of Gaussians of the form

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} \mathcal{N}(o_t; \mu_{jk}, \Sigma_{jk}) \tag{1}$$

where o_t is the d-dimensional observation vector at time frame t, K is the number of mixture components, c_{jk} is the mixture coefficient for the k-th mixture in state j, and (μ_{jk}, Σ_{jk}) are the mean vector and the covariance matrix of the Gaussian k-th component of the j-th state distribution.

The SAT re-estimation process is depicted in Figure 1. The feedback lines indicate that the process can be iterated, until convergence to the optimal point is obtained. Each iteration of SAT consists of two phases, the adaptation-training-estimation (ATE) phase, and the synchronization (SYNC) phase.

In the i-th iteration of SAT, the SI model λ_{i-1} from the prior iteration is adapted to each of the speakers in the training set. For the first iteration ($i = 1$), λ_0 is initialized to a sufficiently trained SI model. During the adaptation phase, the SI means are mapped to the unknown speaker dependent (SD) means by a linear regression transform $G_{i-1}^{(s)} = \left(W^{(s)}, \beta^{(s)}\right)$ as follows

$$\mu_{jk}^{(s)} = W^{(s)} \mu_{jk} + \beta^{(s)} \tag{2}$$

where $W^{(s)}$ is a $d \times d$ transformation matrix and $\beta^{(s)}$ is an

additive bias vector. The index $i-1$ in $G_{i-1}^{(s)}$ indicates that this transformation is estimated from the adaptation data during the prior iteration of SAT, using the Maximum Likelihood Linear Regression (MLLR) method [2]. For the first iteration of SAT, $G_0^{(s)}$ is initialized to the identity transform ($W^{(s)} = I_d$ and $\beta^{(s)} = 0$).

After the transformation of the SI means, the j-th state observation density adapted to speaker s is given by

$$b_j(o_t^{(s)}) = \sum_{k=1}^{K} c_{jk} \mathcal{N}(o_t^{(s)}; W^{(s)}\mu_{jk} + \beta^{(s)}, \Sigma_{jk}) \qquad (3)$$

In what follows, we shall assume that the speaker specific transformation consists of a single regression matrix for simplicity. It is possible, however, to define regression classes and associate a regression matrix with each class. The extension of the SAT parameter estimation to multiple regression transformations is straightforward.

The adaptation of λ_{i-1} to speaker s produces a SD model $\lambda_{i-1}^{(s)}$ which in turn is used as the seed model for training on the speaker data using the forward-backward algorithm [3]. The resulting model $\lambda_i^{(s)}$ together with the original SI model λ_{i-1} are fed forward to the estimation stage, where the transformation $G_i^{(s)}$ is estimated using MLLR. This completes the ATE phase of the SAT process.

The SYNC phase is not executed until models $\lambda_i^{(s)}$ and transformations $G_i^{(s)}$ have been obtained for all the speakers in the training set. Then, the means and variances of the output SI model λ_i are re-estimated as follows

$$\begin{aligned} \overline{\mu}_{jk} &= \left\{ \sum_s^S \gamma_{jk}^{(s)} {W^{(s)}}^T \Sigma_{jk}^{-1} W^{(s)} \right\}^{-1} \times \\ &\quad \left\{ \sum_s^S \gamma_{jk}^{(s)} {W^{(s)}}^T \Sigma_{jk}^{-1} \left(\tilde{\mu}_{jk}^{(s)} - \beta^{(s)} \right) \right\} \end{aligned} \qquad (4)$$

$$\overline{\Sigma}_{jk} = \frac{\sum_s^S \gamma_{jk}^{(s)} \left[\tilde{\Sigma}_{jk}^{(s)} + (\tilde{\mu}_{jk}^{(s)} - \overline{\mu}_{jk}^{(s)})(\tilde{\mu}_{jk}^{(s)} - \overline{\mu}_{jk}^{(s)})^T \right]}{\sum_s^S \gamma_{jk}^{(s)}} \qquad (5)$$

where $\gamma_{jk}^{(s)}$ is the expected number of times the system is in state j using the k-th mixture component [1], $(\tilde{\mu}_{jk}^{(s)}, \tilde{\Sigma}_{jk}^{(s)})$ are the mean and variance of the k-th Gaussian component of the j-th state distribution in model $\lambda_i^{(s)}$, and $\overline{\mu}_{jk}^{(s)} = W^{(s)}\overline{\mu}_{jk} + \beta^{(s)}$ are the Gaussian means adapted to each speaker s using the updated values of the transformation parameters and the Gaussian mean vectors.

[1] $\gamma_{jk}^{(s)}$ is also termed as *mass* of the k-th component of the j-th state distribution for speaker s

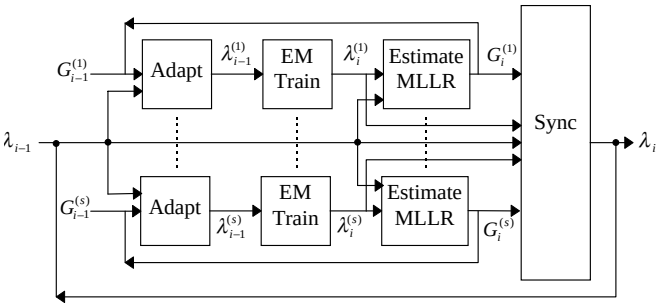

Figure 1: Block diagram of original SAT method

2.1. Computational Requirements

In order to evaluate the computational requirements of the original SAT method, as well as those of the three proposed SAT implementations, we will compare them to the requirements of the common SI training paradigm. In particular, in terms of disk space requirements, any training method requires space for the input (seed) and output acoustic models, and this space is the same for all the methods, therefore we will concentrate on the intermediate required disk space and the associated I/O operations. Additionaly, we shall assume that each real number in all systems is stored in 4 bytes of space.

Disk space. From Equations 4, 5 we see that the original SAT method requires for each speaker s the storage of the Gaussian mean vectors $\tilde{\mu}_{jk}^{(s)}$, the variances $\tilde{\Sigma}_{jk}^{(s)}$, the masses $\gamma_{jk}^{(s)}$, and the transformation $G^{(s)} = (W^{(s)}, \beta^{(s)})$. These are the parameters that the synchronization phase requires in order to re-estimate the means and variances of the SI model. This is a significant requirement of disk space per speaker. Assume that we have N states, with K Gaussian mixture components per state. Then, each speaker model has NK Gaussian d-dimensional mean and variance vectors (variances are diagonal matrices), and NK masses. Hence the space required for the speaker model parameters is $NK(2d+1)$. Also, we need to store the speaker transformations, i.e. an additional $d(d+1)$ per speaker. Therefore the total required space per speaker is $NK(2d+1) + d(d+1)$.

Assuming that there are S speakers in the training set, the total required disk space is $S[NK(2d+1) + d(d+1)]$. As an example, consider training on 2000 speakers using the original SAT method. In our typical State Clustered Tied Mixtures (SCTM) system, $N = 3000$, $K = 64$, and $d = 45$. If each vector component is represented with 4 bytes, then the total required disk space would be 130 GBytes!

Compute time. The overhead that SAT adds on top of the common SI training is concentrated mainly on disk I/O operations that take place in the training stage of the ATE phase and in the SYNC phase. The adaptation and estimation stages, require time proportional to the number of regression classes as well as the amount of adaptation data. Especially the estimation stage can be compute-intense in the case there are many regression classes defined, since each transformation

matrix associated with a regression class needs a series of matrix inversions in order to be estimated from the adaptation data.

It is clear from the above discussion that the excessive requirements of disk space and I/O make the original SAT method impractical for training on real-life conditions which involve thousands of speakers. In the following section we present three alternative implementations that try to overcome this problem.

3. ALTERNATIVE SAT IMPLEMENTATIONS

The practical implementations of SAT presented here focus mainly on reducing the disk space requirement as much as possible. We start with a modest solution, the 2-pass SAT, which basically implements the original SAT theoretical formulation but overcomes the problem of the excessive disk space requirement by executing the training in two passes instead of one. Then, we present two more radical approaches, the Inverse Transform SAT (ITSAT) and the Least Squares SAT (LSSAT), that require minimal disk storage per speaker and try to approximate the original SAT method.

3.1. 2-pass SAT

The main difference of the 2-pass SAT from the original method is that each iteration of SAT requires two passes in order to be completed.

The first pass consists of the ATE phase of the original SAT method and a synchronization phase that re-estimates only the means of the SI model. From Equation 4 we see that the term

$$A_{jk} = \sum_{s}^{S} \gamma_{jk}^{(s)} {W^{(s)}}^T \Sigma_{jk}^{-1} \tilde{\mu}_{jk}^{(s)}$$

can be accumulated in the ATE phase, at the end of the EM training stage for each speaker. A_{jk} is a d-dimensional vector, and it is speaker-independent. The synchronization phase of the first pass reads in the vectors A_{jk} for each state j and mixture component k, as well as the speaker masses $\gamma_{jk}^{(s)}$ and the transformation matrices $W^{(s)}$, and re-estimates the means of the SI model as follows

$$\overline{\mu}_{jk} = \left\{ \sum_{s}^{S} \gamma_{jk}^{(s)} {W^{(s)}}^T \Sigma_{jk}^{-1} W^{(s)} \right\}^{-1} \times A_{jk} \qquad (6)$$

Then the updated mean vectors are stored on the disk, the intermediate parameters A_{jk} and $\gamma_{jk}^{(s)}$ are deleted, and the second pass begins.

The second pass reads the re-estimated means and repeats the adaptation and training steps, but does not need to repeat the estimation step, since the transformation matrices have already been estimated and stored in the first pass. At the end of the EM training stage for a speaker s the parameters $\gamma_{jk}^{(s)}$, $\tilde{\mu}_{jk}^{(s)}$, $\tilde{\Sigma}_{jk}^{(s)}$ and $\overline{\mu}_{jk}^{(s)}$ are available, so we can accumulate the numerator and denominator of the fraction in Equation 5. Therefore, at the end of the second pass we can update the variances of the SI model.

Required disk space. The 2-pass SAT method needs to store the transformations $W^{(s)}$, the speaker masses $\gamma_{jk}^{(s)}$, and the speaker-independent vectors A_{jk}. Assuming again that we have S speakers, N states and K Gaussian mixture components per state, then the total required disk space is $S[d(d+1) + NK] + NKd$. Taking the same example that we used for the original SAT method, ($S = 2000$, $N = 3000$, $K = 64$, $d = 45$), the 2-pass SAT method would require 1.5 GBytes of disk space, which is approximately 87 times less than the original method.

Required compute time. The 2-pass SAT method performs the adaptation and EM training twice, and the estimation of the speaker transforms once. But it requires significantly fewer I/O operations than the original method, which can compensate for the repetition, especially if the I/O subsystem is slow.

Although the 2-pass SAT requires much less disk space than the original method does, it still performs significant amount of disk I/O operations. Also, it repeats the adaptation and training stages in the second pass, so it is still much slower than regular SI training. It would be nice if we could find an SAT re-estimation process that is closer to the computational requirements of SI training, without significant degradation from the performance of the original SAT method. The next two methods try to achieve this goal.

3.2. Inverse Transform SAT

The Inverse Transform SAT (ITSAT) is depicted in Figure 2. The first thing that one can notice from the schematic diagram is the lack of a synchronization stage, which is the main advantage of this method. Each iteration of ITSAT performs exactly the same steps as the ATE phase of the original SAT method, but as soon as the speaker transform has been estimated, it is inverted and applied to the means and variances of the speaker model $\lambda_i^{(s)}$, producing the model $\hat{\lambda}_i^{(s)}$. The transformed means and variances are accumulated over all the speakers in the training, producing the SI model λ_i.

In particular, let $\tilde{\mu}_{jk}^{(s)}$, $\tilde{\Sigma}_{jk}^{(s)}$ be the mean and variance of the k-th Gaussian component of the j-th state in model $\lambda_i^{(s)}$, which is produced at the end of the EM training stage for speaker s. We use MLLR again to estimate the transform $G_i^{(s)} = (W^{(s)}, \beta^{(s)})$ from the SI model λ_{i-1} to the SD model $\lambda_i^{(s)}$. Then we compute an inverse transform ${G_i^{(s)}}^{-1} = (\hat{W}^{(s)}, \hat{\beta}^{(s)})$, from the SD model to the SI model [2], and we apply it to the means and variances as follows

[2] The inversion procedure is not straightforward, because for some speakers the transformation matrices $W^{(s)}$ may be ill-conditioned. We will describe the procedure in detail in the next subsection.

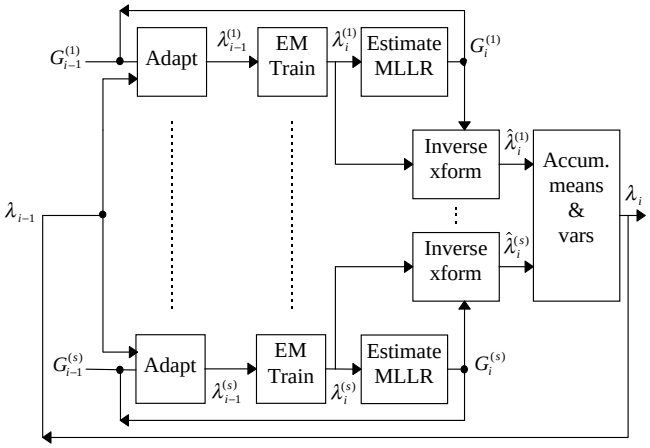

Figure 2: Block diagram of ITSAT method

$$\hat{\mu}_{jk}^{(s)} = \hat{W}^{(s)}\tilde{\mu}_{jk}^{(s)} + \hat{\beta}^{(s)} \tag{7}$$

$$\hat{\Sigma}_{jk}^{(s)} = \hat{W}^{(s)}\tilde{\Sigma}_{jk}^{(s)}\hat{W}^{(s)^T} \tag{8}$$

where $\hat{\mu}_{jk}^{(s)}$, $\hat{\Sigma}_{jk}^{(s)}$ denote the transformed mean and variance, respectively, of the k-th Gaussian component of the j-th state distribution.

The transformed means and variances are accumulated and the SI model parameters are re-estimated as follows

$$\overline{\mu}_{jk} = \frac{\sum_s^S \gamma_{jk}^{(s)}\hat{\mu}_{jk}^{(s)}}{\sum_s^S \gamma_{jk}^{(s)}} \tag{9}$$

$$\overline{\Sigma}_{jk} = \frac{\sum_s^S \gamma_{jk}^{(s)}[\hat{\Sigma}_{jk}^{(s)} + (\hat{\mu}_{jk}^{(s)} - \overline{\mu}_{jk})(\hat{\mu}_{jk}^{(s)} - \overline{\mu}_{jk})^T]}{\sum_s^S \gamma_{jk}^{(s)}} \tag{10}$$

Inversion of transform. In order to compute the inverse transform $G_i^{(s)^{-1}}$ we need to invert the matrix $W^{(s)}$. Experiments showed that $W^{(s)}$ may be ill conditioned for some speakers, so even small roundoff errors that can occur during the inversion of the matrix can have a drastic effect on the computed inverse, and consequently, on the transformed means $\hat{\mu}_{jk}^{(s)}$ and variances $\hat{\Sigma}_{jk}^{(s)}$. This problem can be overcome by "smoothing" the matrix $W^{(s)}$ before computing the inverse. In particular, $W^{(s)}$ can be interpolated with the $d \times d$ identity matrix I_d to obtain a smoothed matrix $\tilde{W}^{(s)}$ as follows

$$\tilde{W}^{(s)} = \alpha I_d + (1-\alpha)W^{(s)} \tag{11}$$

where $\alpha \in [0,1]$ is a parameter that depends on the conditioning of $W^{(s)}$ (it is an increasing function of the conditioning of $W^{(s)}$ in $[0,1]$).

Now the smoothed matrix $\tilde{W}^{(s)}$ can be inverted, but it is not consistent with the additive vector $\beta^{(s)}$ of the original transformation. In our implementation of the ITSAT method, we compute the adjusted vector $\tilde{\beta}$ from the following equation

$$\sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}^{(s)}(W^{(s)}\mu_{jk} + \beta^{(s)}) = \sum_{j=1}^{N}\sum_{k=1}^{K}\gamma_{jk}^{(s)}(\tilde{W}^{(s)}\mu_{jk} + \tilde{\beta}^{(s)}) \tag{12}$$

where μ_{jk} is the mean of the k-th Gaussian component of the j-th state distribution in the SI model λ_{i-1}. In other words, the additive vector is adjusted so that the average mean of the SI model is mapped to the same position in the SD space with both the original and smoothed transformations.

After the smoothed transform $\tilde{G}_i^{(s)} = (\tilde{W}^{(s)}, \tilde{\beta}^{(s)})$ has been computed, we can invert it to obtain

$$\tilde{G}_i^{(s)^{-1}} = (\tilde{W}^{(s)^{-1}}, -\tilde{W}^{(s)^{-1}}\tilde{\beta}^{(s)}) \tag{13}$$

Required disk space. Since the ITSAT method doesn't have a synchronization phase, it needs to save only the (smoothed) speaker transformations $\tilde{G}_i^{(s)}$ for each speaker s in the training, so that they can be used in the adaptation stage of the next iteration. Thus, if S is the number of speakers in the training set, the required disk space is $Sd(d+1)$. To compare to the other two implementations discussed previously, for $S = 2000$ and $d = 45$, ITSAT would require only 16 MBytes of disk space. This is 8,320 times less than the original method!

Required compute time. ITSAT can be almost as fast as regular SI training, depending on the number of transformations that we have to estimate after the EM training stage for each speaker. If few regression classes have been defined, then the overhead of the estimation and adaptation stages is small compared to the actual training time, and there is essentially no additional disk I/O overhead.

As we will see in section 4, the ITSAT method performs nearly as well as the original SAT method, inspite of its approximation to the original SAT formulation. One drawback of this method is that it requires tuning of the parameter α in Equation 11. Our experiments showed that the choice of the function that relates α to the conditioning of $W^{(s)}$ plays a very important role in the success of the method, and there is no rule for choosing such a function. Also, we noticed in our experiments that the resulting SI model λ_i exhibits a variance irregularity: in the original and 2-pass SAT methods, λ_i is a *reduced-variance model*, i.e., its average variance is smaller than the one in λ_{i-1} over all dimensions; in ITSAT however, the average variance of λ_i is larger than the average variance of λ_{i-1} in some dimensions and smaller in others.

The following implementation tries to solve the above problems by *estimating* the inverse transform from the speaker data using the least squares method, which guarantees a re-

duced variance SI model.

3.3. Least Squares SAT

The schematic diagram of the Least Squares SAT (LSSAT) method is shown in Figure 3. We can see that LSSAT is similar to the ITSAT method in that it doesn't require a synchronization phase. But unlike ITSAT, LSSAT doesn't have an adaptation stage.

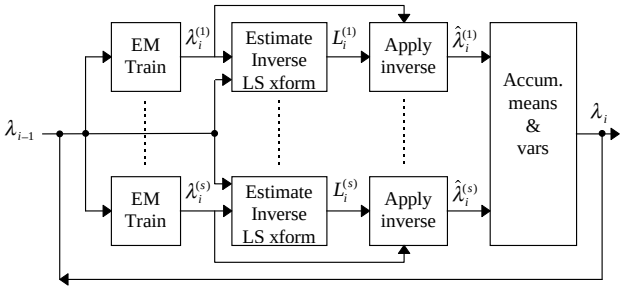

Figure 3: Block diagram of LSSAT method

In the i-th iteration of LSSAT, the SI model λ_{i-1} from the prior iteration is used as a seed model for training on the utterances of speaker s using the forward-backward algorithm. The resulting model $\lambda_i^{(s)}$ together with the original SI model λ_{i-1} are then fed forward to the estimation stage, where the inverse transform $L_i^{(s)} = (\hat{W}^{(s)}, \hat{\beta}^{(s)})$ is estimated using the Least Squares method.

Then $L_i^{(s)}$ is applied to the means and variances of the SD model as in Equations (7, 8), and the transformed means $\hat{\mu}_{jk}^{(s)}$ and variances $\hat{\Sigma}_{jk}^{(s)}$ are accumulated over all speakers in the training set to re-estimate the SI means and variances as in Equations (9, 10).

Required disk space. Since the LSSAT method doesn't have an adaptation stage, it doesn't need to store any speaker transformation matrices. Therefore there is no additional disk space required by the method besides the space for the input and output models λ_{i-1} and λ_i (which is the same as the space required in the common SI training paradigm).

Required compute time. LSSAT is the fastest of the three proposed implementations of SAT, and the only overhead that it adds on top of the regular SI training is the estimation of the inverse transform $L_i^{(s)}$ and its application to the SD means and variances. It is important to note here that the least squares estimation method is much faster than the MLLR method, since it requires only one matrix inversion per regression class in order to compute the inverse transformation matrix. Therefore, the computational requirements of LSSAT match approximately those of the common SI training paradigm.

Although the LSSAT method requires no additional disk space and has minimal overhead compared to the SI training, it doesn't perform as well as the original SAT method as we will see in the next section.

4. EXPERIMENTAL RESULTS

In this section we present the results of recognition experiments that we conducted in order to evaluate the efficacy of the three proposed implementations of SAT. These experiments use BYBLOS, BBN's state-of-the-art large vocabulary recognition system [4]. As the baseline speaker-independent system for theses experiments we use a gender dependent triphone-based continuous density HMM system. All allophone models of each of the 46 phonemes of the system are modeled by a mixture density of 256 Gaussian components in a configuration termed as Phonetically Tied Mixture (PTM) HMM. Speech is parameterized using 14 mel-warped cepstral coefficients, a short-term power coefficient and the first and second order difference of these parameters to give a 45 dimensional feature vector.

To simplify experimentation, and reduce the turnaround time, we used acoustic training data consisting of 10 hours of speech, collected from 200 male speakers from the Hub-4 1996 Broadcast News (BN) corpus. Following the common SI paradigm, we constructed a SI male acoustic model, to use as the initial seed model for all the SAT re-estimation procedures outlined in the previous sections. We evaluated the efficacy of the three new SAT implementations by comparing the recognition performance of their models to that of the original SAT acoustic model, with unsupervised adaptation to the test speakers. The adaptation to the test procedure is the same for all acoustic models, either SI or SAT, and it uses the MLLR method, with 2 regression classes defined per speaker. The testing material consists of the male speakers in episodes k,l,n of the Hub-4 1996 BN development test. The total number of words in these three episodes (male speakers only) is 8611.

Acoustic model	Word Error %
SI unadapted	29.5
SI adapted	27.5
Original SAT	26.4
2-pass SAT	26.4
ITSAT	26.7
LSSAT	27.6

Table 1: Word Error Rate (%) comparisons

The results are shown in Table 1, where we have also included the recognition results for the SI acoustic models, both unadapted and adapted to the test speakers, to see the overall improvement that we obtain with the SAT implementations. We can see that the 2-pass SAT method is identical to the original SAT method. This was expected, since the two methods share the same theoretical formulation. Also, the result of the ITSAT method is very close to that of the original SAT method; it has only 1% degradation, and we believe that this result can be improved by a better choice of the smoothing function. Finally, the LSSAT method has the worst result between the three practical SAT implementations, and it is not as good as the SI adapted model result.

5. CONCLUSIONS

We have presented three implementations of SAT that are practical for training on thousands of speakers without requiring speaker clustering. The 2-pass SAT method implements the original SAT theoretical formulation, but performs each iteration in two passes. It achieves the same recognition performance as the original SAT method with much less requirements in disk space and disk I/O operations. The ITSAT method, although much simpler and with computational requirements that approach those of common SI training, performs as good as the original SAT method, with only 1% degradation. Finally, we presented the LSSAT implementation that uses the least squares method to estimate the inverse transform from the speaker data. LSSAT was motivated by the difficulty of inverting the MLLR transformation in ITSAT, but the recognition results showed that its performance is not as good as that of the SI adapted model.

Clearly, ITSAT is the most attractive of the three proposed SAT implementations. We believe that ITSAT can achieve even better recognition performance by tuning the interpolation parameter α to give adequate smoothing for ill conditioned transformation matrices.

Acknowledgments

This work was supported by the Advanced Research Projects Agency and monitored by Ft. Huachuca under contract No. DABT63-94-C-0063. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the Government and no official endorsement should be inferred.

References

1. T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul "A Compact Model for Speaker-Adaptive Training", *ICSLP Proceedings*, October 3-6 1996 Philadelphia, PA.
2. C.J. Leggetter and P.C. Woodland, "Speaker Adaptation of HMMs Using Linear Regression", Tech. Rep. CUED/FINFENG/TR.181, Cambridge University Engineering Department, June 1994.
3. L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", *Ann. Math. Stat.*, vol. 41, pp. 164-171, 1970.
4. F. Kubala, H. Jin, S. Matsoukas, L, Nguyen, R. Schwartz, J. Makhoul, "The 1996 BBN Byblos Hub-4 Transcription System", elsewhere in this volume.